Thomas Burch

Fragen der Vernetzung in OWID[1] und im *Wörterbuchnetz*[2]

Abstract

Nachschlagewerke sind aufgrund ihres primären Verwendungszweckes, Informationen schnell und gezielt zu finden, auf vielfältige Art aufeinander bezogen und damit bereits in der gedruckten Fassung in gewisser Weise sowohl explizit als auch implizit vernetzt angelegt. Im Falle elektronischer Nachschlagewerke lassen sich diese „Netzwerke" zusätzlich ausweiten, indem auch Beziehungen zwischen Informationseinheiten etabliert werden können, die in der gedruckten Fassung, beispielsweise aufgrund der Entstehungsgeschichte der Einzelwerke, bisher gar nicht möglich waren. Diese Vernetzungen können weit über eine rein ausdrucksseitige Verknüpfung hinaus, indem sie philologische und informationswissenschaftliche Methoden verbinden. Im Folgenden werden die Wörterbuchverbünde *OWID* und *Wörterbuchnetz* vorgestellt und insbesondere auf das darin enthaltene Vernetzungspotential eingegangen. Neben den in beiden Ansätzen vorhandenen expliziten Verweise zwischen den Wortartikeln werden für das Trierer Wörterbuchnetz zusätzlich automatische Verfahren und Methoden aufgezeigt, mit deren Hilfe bisher nur implizit gegebene Beziehungen zwischen den Wortartikeln ermittelt und zur Überprüfung vorgeschlagen werden können.

Because of their primary usage, i.e. to retrieve information very fast and systematically, dictionaries are strongly cross-linked. Already in the printed edition this network is established in an explicit and an implicit way. But in the case of an electronic publication these interconnections can be additionally extended by integrating relations between information positions, which until now could not be part of the printed versions. The reason for these missing links results e.g. from the historical order of dictionary development, older ones can not link to newer ones. But these connections can not only be found by simply considering similarities within the text patterns. They can be calculated with the help of algorithms from the field of information retrieval and statistical analysis. In the following the two dictionary networks *OWID* and *Wörterbuchnetz* will be presented especially focussing on their capabilities of cross-linking. Beside the basic concepts of the two approaches we will demonstrate the underlying methods and algorithms for the *Wörterbuchnetz* by the help of which we are able to calculate and measure similarities between the different dictionaries.

Digitale Nachschlagewerke sind, wie auch ihre gedruckten Entsprechungen, auf vielfältige Art aufeinander bezogen und damit in gewisser Weise implizit „vernetzt"; eine übergreifende, integrierte Recherche ist jedoch wegen der Unterschiede in der Anlage, Anordnung und Struktur der einzelnen Werke auch im digitalen Medium nicht ohne Weiteres möglich.[3] Nachschlagewerke, die durch inhaltlich-strukturelles Markup in standardisierte und damit vergleichbar gemachte Informationseinheiten gegliedert und durch Metadaten angereichert sind, machen jedoch die impliziten Vernetzungen explizit. Dadurch kann eine neue Qualität der Informationsgewinnung erreicht und die Lücke zwischen der schwerfälligen Benutzbarkeit und eingeschränkten Verfügbarkeit der Buchversionen einerseits und der fehlenden Systematik und Beliebigkeit der Information im Internet andererseits geschlossen werden.

[1] *OWID* (*Online-Wortschatz-Informationssystem Deutsch*), das Portal für lexikografische Arbeiten am Institut für Deutsche Sprache (IDS), Mannheim, www.owid.de (zuletzt eingesehen am 27.12.2007). Im Folgenden wird OWID (bzw. *elexiko*) im Stand von Mai 2007 präsentiert, da sich am hier Gezeigten inhaltlich nichts Grundsätzliches verändert hat.

[2] *Das Wörterbuchnetz,* Verbund von neun digitalisierten Wörterbüchern, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier, www.woerterbuchnetz.de (zuletzt eingesehen am 27.12.2007).

[3] Diese unterschiedlichen Strukturen erklären sich zum einen wissenschaftshistorisch aus der Entwicklung der lexikografisch-lexikologischen Methoden, zum anderen aus den spezifischen Zielsetzungen der verschiedenen Wörterbuchtypen.

Im Gegensatz zu einer bloßen, unverbundenen Bereitstellung verschiedener Nachschlagewerke bieten Wörterbuchverbünde mit multidirektionalen Verlinkungen komplexe und gezielt spezifizierbare Zugänge zum Material. Diese Verlinkungen gehen weit über eine rein ausdrucksseitige Verknüpfung hinaus, indem sie philologische und informationswissenschaftliche Methoden verbinden. Die Entwicklung von Verfahren, Methoden und Technologien, die eine intelligente, dynamische Vernetzung von Nachschlagewerken unterschiedlichen Typs einschließlich ihrer Quellenverzeichnisse und Primärquellen ermöglichen und dadurch eine neue Qualität des Informations- und Wissensmanagements erreichen, erfordert neben informationstechnologischen Konzepten auch eine stärkere Einbeziehung philologischer und lexikografischer Verfahrensweisen. Im Folgenden werden die Wörterbuchverbünde OWID und *Wörterbuchnetz* vorgestellt und insbesondere wird auf das darin enthaltene Vernetzungspotenzial eingegangen. Neben den in beiden Ansätzen vorhandenen expliziten Verweisen zwischen den Wortartikeln werden für das Trierer Wörterbuchnetz zusätzlich automatische Verfahren und Methoden aufgezeigt, mit deren Hilfe bisher nur implizit gegebene Beziehungen zwischen den Wortartikeln ermittelt und zur Überprüfung vorgeschlagen werden können (vgl. Rapp 2006).

In OWID, dem Informationssystem zum deutschen Wortschatz des Instituts für Deutsche Sprache (IDS) (vgl. Klosa et al. 2006), werden die Forschungsergebnisse verschiedener Projekte virtuell vereinigt. Die korpusbasiert erarbeiteten Nachschlagewerke werden unter einer einheitlichen grafischen Benutzeroberfläche elektronisch im Internet publiziert. Zurzeit umfasst es die Komponenten *elexiko*, *Neologismenwörterbuch*, *Wortverbindungen online* und ein Nachschlagewerk zum *Schulddiskurs im ersten Nachkriegsjahrzehnt*. Das *elexiko*-Wörterbuch ist auf über 300.000 Stichwörter ausgelegt, von denen zurzeit etwa 800 ausführlich lexikografisch beschrieben sind. Das *Neologismenwörterbuch* präsentiert in mehr als 750 umfangreichen Wortartikeln neue Wörter, neue feste Wortverbindungen sowie neue Bedeutungen von etablierten Wörtern, die in den 90er-Jahren des 20. Jahrhunderts in die Allgemeinsprache eingegangen sind. Im Modul *Wortverbindungen online* werden empirische Ergebnisse auf dem Gebiet der korpusgesteuerten Mehrwortforschung veröffentlicht, die sich für eine Online-Präsentation eignen. Die Präsentationsformen haben unterschiedliche linguistische Beschreibungstiefen und Darstellungsformate. Derzeit sind ca. 130 lexikografische Mehrwortartikel über die produktinterne Stichwortliste abrufbar. Das Wörterbuch zum *Schulddiskurs im ersten Nachkriegsjahrzehnt* verzeichnet 85 Haupt- und über 200 Unterstichwörter. Jeder Artikel ist mit einem umfangreichen Beleganhang versehen. Der hier dargestellte Wortschatzbereich ist erarbeitet worden aus einem breit angelegten Korpus von Texten, die in den Jahren 1945 bis 1955 erschienen sind.

Die inhaltlichen Bezüge zwischen den einzelnen Modulen bieten dabei ein großes Vernetzungspotenzial, welches durch das elektronische Medium in idealer Weise für den Benutzer aufbereitet werden kann. Aufgrund der dynamischen Weiterentwicklung des Artikelbestandes ist es jeder gedruckten Publikation weit überlegen.

Allerdings müssen dazu entsprechende Voraussetzungen in Form von wohldefinierten Konzepten geschaffen werden, damit ein elektronisches System einwandfrei funktionieren kann. Dies beginnt in der Regel mit einer standardisierten Datencodierung, die eine konzeptuelle Inhaltsmodellierung der Artikelstrukturen beinhaltet. In OWID wird hier eine einheitliche XML-Codierung zugrunde gelegt (vgl. Müller-Spitzer 2007a), die (zusammen mit dem geplanten Einsatz von Xlink/XPointer (vgl. Müller-Spitzer 2007b))

einen plattform- und systemneutralen Datenbestand garantiert, in dem sowohl die inhaltlichen wie auch die Vernetzungskonzepte enthalten sind. Aufgrund dieser konsequenten Datencodierung zeigt sich das Vernetzungspotenzial in OWID, für den Benutzer völlig transparent, auf unterschiedlichen Ebenen. Eine den Modulen übergeordnete Vernetzung ergibt sich durch die gemeinsame Stichwortliste, in der alle behandelten Artikel aufgeführt werden. Durch Auswahl eines Eintrags wird der Benutzer in das zugehörige Teilmodul geführt (vgl. Abbildung 1).

In dem gezeigten Beispiel werden aus der Stichwortliste von *elexiko* auch Einträge in den Modulen zum *Schulddiskurs im ersten Nachkriegsjahrzehnt* bzw. in den *Wortverbindungen online* aufgerufen. Verbunden mit der übergeordneten Stichwortsuche kann hier also über eine gemeinsame Schnittstelle der gesamte Artikelbestand abgefragt werden.

Eine weitere Vernetzungsebene ergibt sich durch Bezüge im Innern der Wortartikel, indem Referenzen auf verwandte Stichwörter durch entsprechende Verweise realisiert werden. Dabei spielt es keine Rolle, ob diese Verweise zwischen den Modulen wie im Falle der Stichwortliste oder innerhalb eines einzelnen Moduls bestehen. Eine solche interne Modulvernetzung zeigt Abbildung 2.

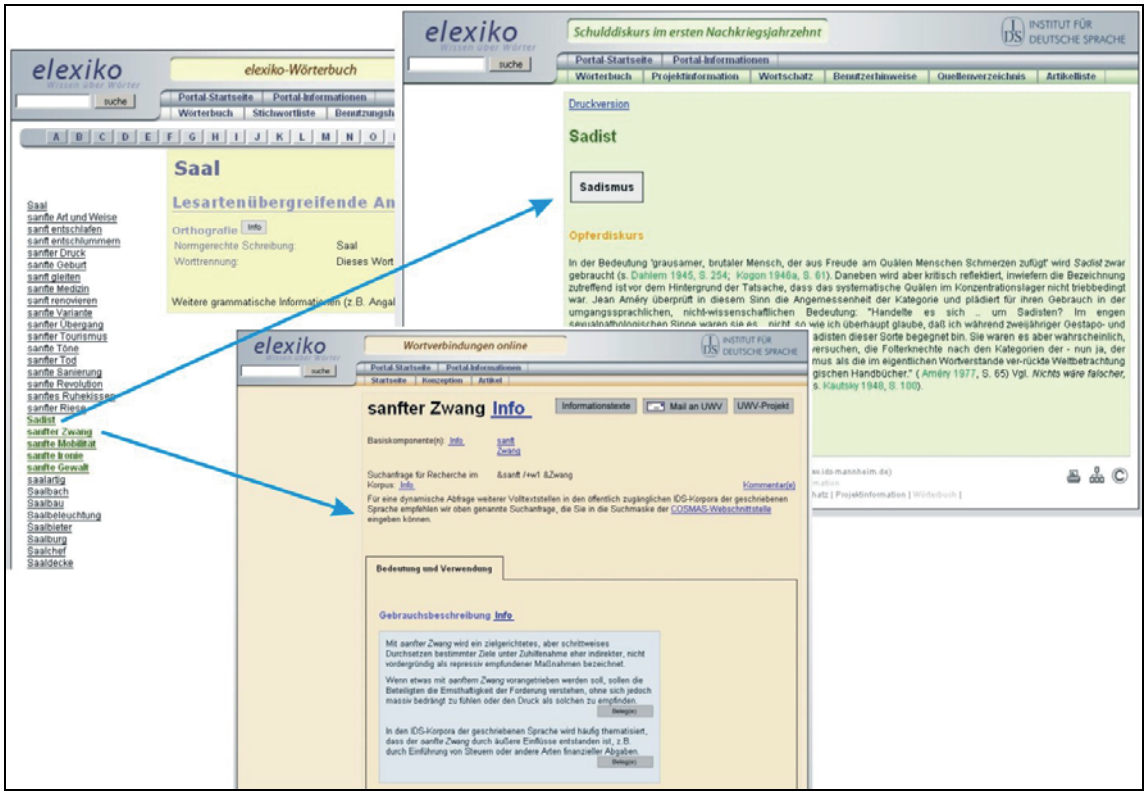

Abbildung 1: Vernetzung zwischen Einzelmodulen in OWID

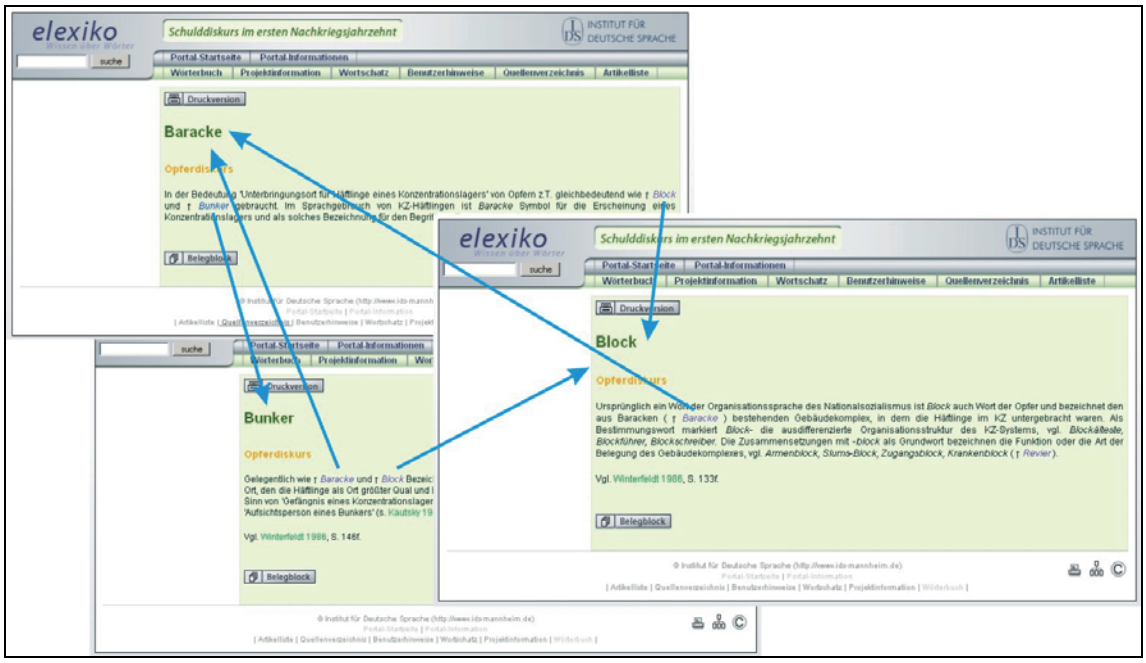

Abbildung 2: Vernetzung innerhalb eines Wörterbuches in OWID

In dem gezeigten Beispiel verweisen die Nennungen verwandter Stichwörter im Artikeltext auf die zugehörigen Einträge in OWID. Prinzipiell sind diese Verweise an jeder Stelle des Artikeltextes möglich. Es ergeben sich dadurch auch Verweise unterschiedlicher Qualität, indem beispielsweise synonyme oder antonyme Bezüge, aber auch Relationen zwischen Ober- und Unterbegriff dargestellt werden können. Wie im Beispiel zu sehen, können die Verweise bidirektional eingerichtet sein und durch sukzessive Verkettung Stichwörter zu Wortgruppen zusammengefasst werden. Die Begriffe *Baracke*, *Block* und Bunker sind durch die transitive Vernetzung implizit zu einem Wortcluster kombiniert. Eine Erweiterung der grafischen Oberfläche von OWID könnte darin bestehen, diese bisher nur intern gegebenen Strukturen zu visualisieren. Da das gesamte Material dynamisch erweitert wird, würde dieses Netzwerk mit jedem neu aufgenommenen Artikel dichter werden. Durch die redaktionell betreute Artikelarbeit kann hier eine sehr hohe Qualität der Vernetzungsinformation garantiert werden. Der geplante Einsatz von XLink/XPointer soll darüber hinaus zusätzliche Verbesserung bewirken, z.B. im Bereich der Typisierung der Links, der differenzierteren Visualisierbarkeit, aber auch in der Verwaltung der Vernetzungen für die Lexikografen.

Gegenüber dem dynamisch wachsenden Datenbestand in OWID stellt sich die Ausgangslage für das Trierer Wörterbuchnetz anders dar. Hier stehen unterschiedliche Typen tief annotierter Wörterbücher in digitaler Form zur Verfügung, die bereits in unterschiedlicher Dichte untereinander vernetzt, durch Symptomwertangaben und Metadaten intellektuell erschlossen und recherchierbar oder auch mit Primärquellen vernetzt sind:

- Typ Sprachstadienwörterbuch: Mittelhochdeutsches Wörterbuch, Mittelhochdeutsches Handwörterbuch (einschlägige Nachträge), Findebuch zum mittelhochdeutschen Wortschatz (als Verbund mit multidirektionalen Verweisen), Neues Mittelhochdeutsches Wörterbuch einschließlich digitalem Quellen- und Belegarchiv

- Typ sprachstadienübergreifendes Wörterbuch: Deutsches Wörterbuch von Jacob und Wilhelm Grimm (DWB)
- Typ Dialektwörterbuch: Pfälzisches Wörterbuch, Rheinisches Wörterbuch, Wörterbuch der elsässischen Mundarten, Wörterbuch der deutsch-lothringischen Mundarten, Luxemburgische Wörterbücher
- Typ Autorenwörterbuch: Goethe-Wörterbuch (GWB)
- Typ Sachwörterbuch: Ökonomische Enzyklopädie von J. G. Krünitz (Projekt der UB Trier)

Ziel der Überlegungen ist, anhand der verschiedenen Wörterbücher zum „Deutschen" einen standardisierten Einstieg zu schaffen, von dem aus sich die nicht standardisierten, heterogen organisierten und verstreuten Informationen verzweigen – heterogen im Hinblick auf synchrone wie diachrone Varianz (vgl. Burch/Rapp 2007). Dieses Konzept lässt sich nicht nur auf Wörterbücher anwenden, sondern auch auf die Erschließung großer Korpora übertragen.

Geht man, wie im Falle des *Wörterbuchnetzes*, von retrospektiv digitalisierten Wörterbüchern aus, so ist der Aufbau einer entsprechenden Verweisstruktur ungleich schwieriger. Schon aufgrund der riesigen Datenmenge ist eine manuelle Erarbeitung und Prüfung der Verweise kaum in vertretbarem Zeitaufwand zu bewältigen. In der Regel beschränkt sich dies auf die Codierung der explizit im Druck genannten Verweise, die in vielen Fällen durch eindeutige Texteinleitungen (z.B. *s.*, *vgl.*, *s.u.*, *s.v.a.*) zumindest durch eine Programmroutine aufgefunden und dann durch einen Bearbeiter verifiziert werden können. Eine darüber hinausgehende inhaltliche Vernetzung der Wörterbucheinträge ist nur durch den Einsatz automatischer Verfahren denkbar. Benutzbar werden die von diesen Verfahren gelieferten Ergebnisse aber nur dann, wenn sie durch eine anschließende redaktionelle Bearbeitung überprüft und damit in ihrer Qualität kontrolliert werden.

Aus der Sicht der Informatik stellt sich somit zunächst die Frage, mit welchen Methoden sich der Aufbau eines Wörterbuchverbundes algorithmisch unterstützen lässt. Eine erste Ansatzmöglichkeit, die im Folgenden näher vorgestellt werden soll, stammt hier aus dem Bereich des Information-Retrieval. In einem ersten Schritt wird von der Realität abstrahiert und ein geeignetes Modell für einen Wörterbuchverbund erstellt, welches dann als Ausgangspunkt zur Implementierung der Algorithmen dient. Ganz allgemein entspricht einem beliebigen Verbund dabei die informationstheoretische Datenstruktur eines aus Knoten und Kanten bestehenden Graphen (vgl. Mehlhorn 1984). Die Knoten des Graphen repräsentieren zunächst die Wörterbücher und die Kanten beschreiben die Verbindungen, das heißt die Verweise, zwischen den Wörterbüchern. Diese grobe Struktur lässt sich weiter verfeinern und genauer modellieren, indem nicht das gesamte Wörterbuch als Knoten aufgefasst wird, sondern die einzelnen Wortartikel. Die Kanten verlaufen dann zwischen den Artikeln verschiedener Wörterbücher bzw. auch innerhalb eines einzelnen Wörterbuchs. Die Kanten sind gerichtet, das heißt, Ausgangs- und Zielpunkt sind eindeutig bestimmt.

Ausgehend von den gedruckten Wörterbüchern lassen sich die Verweise zunächst in drei Klassen einteilen: 1. Verweise, die explizit im Wörterbuch genannt sind, beispielsweise durch Angabe einer Seiten-, Spalten-, Zeilen- oder Artikelreferenz; 2. Verweise, die sich aufgrund von statistischen Berechnungen der Wörter in den Artikeln identifizie-

ren lassen; 3. Verweise, die sich durch Graphalgorithmen aus den Klassen 1 und 2 berechnen lassen, beispielsweise durch transitiven Abschluss. Mit den im Folgenden vorgestellten Verfahren sollen zunächst Verweise der Klasse 2 bestimmt werden.

Die Ausgangsbasis zur Implementierung der Algorithmen bildet ein Datenbanksystem (siehe Abbildung 3), in welchem die Wörterbücher in strukturierter Form abgespeichert sind. Die Architektur des Gesamtsystems basiert auf einer standardisierten SGML/XML-Codierung (vgl. Goldfarb 1990) der Wörterbuchdaten, die sich nach unterschiedlichen Document-Type-Definitionen (DTDs) richten kann. Zu jeder der vorgegebenen Codierungen existiert ein zugehöriger Importfilter, über den die Wörterbuchdaten in ein Datenbankmanagementsystem (DBMS) (vgl. Rob/Carlos 1993) übernommen werden können. In umgekehrter Richtung wird ein Exportfilter entwickelt, über den die durch die Bearbeitung innerhalb des Systems neu eingefügten Informationen in Form von zusätzlichen Metadaten zu den grundlegenden Wörterbuchdaten ebenfalls in standardisiertem XML-Format aus dem Datenbanksystem herausgezogen werden können. Auf diese Weise stehen sowohl die Wörterbuchdaten als auch die Beziehungen zwischen den Wörterbüchern in einer plattformunabhängigen und damit langfristig nutzbaren Form zur Verfügung.

Das Kernstück des gesamten Systems bildet ein relationales Datenbankmanagementsystem, in dem jeweils ein Wörterbuch durch eine zugehörige Datenbank verwaltet wird. Der Aufbau der einzelnen Datenbanken richtet sich dabei nach der Granularität der Auszeichnung. Als kleinste gemeinsame Schnittmenge wird nur festgelegt, dass separate Tabellen zur Abfrage der Stichwörter existieren. Weitere Konstituenten der Wörterbuchartikel wie beispielsweise Angaben zur Wortart, Bedeutungserläuterungen, Belege, Angaben zur Etymologie usw. werden dann in eigenen Datentabellen verwaltet, wenn sie durch die XML-Auszeichnung innerhalb des Gesamtartikels markiert wurden. Je feiner die Auszeichnung ist, desto differenzierter sind die Suchmöglichkeiten und desto genauer die Suchergebnisse.[4]

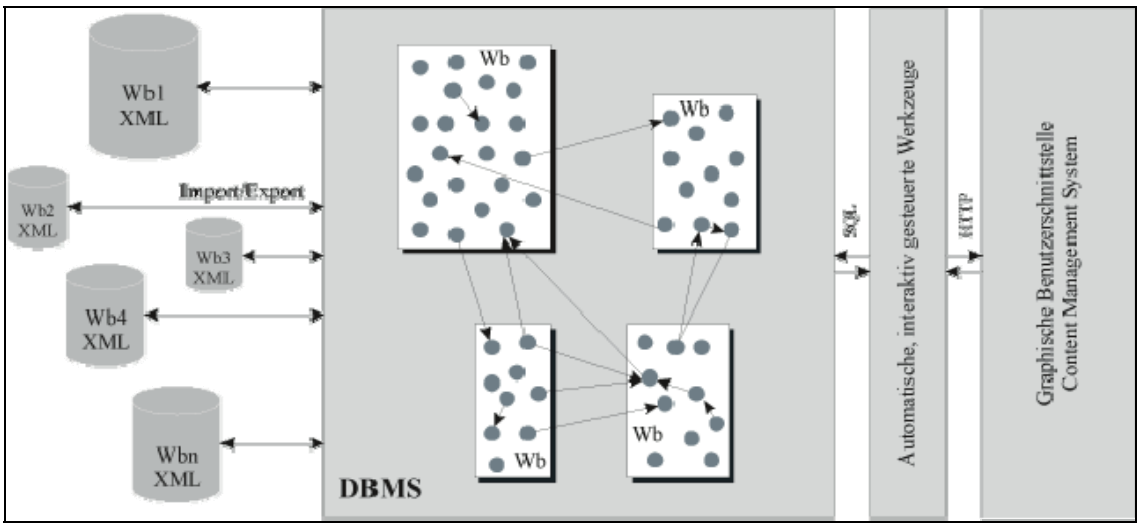

Abbildung 3: Architektur des Datenbanksystems zur Speicherung der Wörterbuchdaten

[4] Bei Wörterbüchern, bei denen die Artikelmikrostrukturen kaum differenziert werden, bliebe im ungünstigsten Fall nur der Zugriff über eine herkömmliche Volltextrecherche.

Ausgehend von den in den Datenbanken abgespeicherten Wörterbuchartikeln können nun automatische Verfahren entwickelt werden, die zunächst zur Aufgabe haben, sämtliche Wortformen der Wörterbücher zu gewichten. Diese Gewichtung erfolgt durch quantitative Angaben über die Häufigkeit einer Wortform (= Term) innerhalb eines Wortartikels und innerhalb des gesamten Wörterbuchs. Berechnet werden die so genannte relative Häufigkeit und die inverse Dokumentfrequenz für jeden Term. Aus beiden Werten ergibt sich dann durch Multiplikation das sogenannte Termgewicht. Ein Beispiel für diese Berechnungsvorlage zeigt Abbildung 4.

TextID	Wort	Normierte Form	ArtikelID	Häufigkeit der Wortform	Wortanzahl im Artikel	Dokumenten-häufigkeit	Inverse Dokumentfrequenz	Gewicht
1507749	butterblume	butterblume	GB13619	1	31	21	14.791	-58.486
1507750	f.	f	GB13619	1	31	93291	2.674	-10.572
1507751	nnl.	nnl	GB13619	1	31	3829	7.280	-28.788
1507752	boterbloem,	boterbloem	GB13619	1	31	1	19.183	-75.854
1507753	gilt	gilt	GB13619	1	31	2896	7.683	-30.381
1507754	für	fuer	GB13619	4	31	40359	3.882	-10.220
1507755	mehrere	mehrere	GB13619	1	31	1408	8.724	-34.495
1507756	kräuter,	kraeuter	GB13619	1	31	467	10.316	-40.791
1507757	für	fuer	GB13619	4	31	40359	3.882	-10.220
1507758	caltha	caltha	GB13619	1	31	31	14.229	-56.264
1507759	palustris,	palustris	GB13619	1	31	163	11.834	-46.795
1507760	dotterblume,	dotterblume	GB13619	1	31	10	15.861	-62.718
1507761	für	fuer	GB13619	4	31	40359	3.882	-10.220
1507762	leontodon	leontodon	GB13619	1	31	42	13.791	-54.531
1507763	taraxacum,	taraxacum	GB13619	1	31	53	13.455	-53.204
1507764	für	fuer	GB13619	4	31	40359	3.882	-10.220
1507765	ranunculus	ranunculus	GB13619	1	31	128	12.183	-48.174
1507766	auricomus.	auricomus	GB13619	1	31	8	16.183	-63.991
1507767	das	das	GB13619	1	31	84491	2.817	-11.137
1507768	volk	volk	GB13619	1	31	4258	7.127	-28.182
TextID	Wort	Normierte Form	ArtikelID	Häufigkeit der Wortform	Wortanzahl im Artikel	Dokumenten-häufigkeit	Inverse Dokumentfrequenz	Gewicht
1507769	wähnt,	waehnt	GB13619	1	31	74	12.974	-51.300
1507770	davon,	davon	GB13619	1	31	5680	6.711	-26.538
1507771	dasz	dasz	GB13619	1	31	35412	4.071	-16.098
1507772	sie	sie	GB13619	1	31	45488	3.710	-14.670
1507773	die	die	GB13619	2	31	123292	2.271	-7.653
1507774	kühe	kuehe	GB13619	1	31	374	10.636	-42.058

Abbildung 4: Berechnung der inversen Dokumentfrequenz für einen Wortartikel aus dem Deutschen Wörterbuch

Die Tabelle zeigt die durch den Algorithmus berechneten Werte für den Artikel *Butterblume* aus dem Deutschen Wörterbuch von Jacob und Wilhelm Grimm. Entscheidend zur Berechnung des Gewichtes (letzte Spalte) sind die Angaben über die Häufigkeit der Wortform, die Länge des Artikels (gemessen in Wortformen) sowie die Dokumentenhäufigkeit, die die Anzahl der Dokumente (= Wortartikel) repräsentiert, in denen der betreffende Term auftritt. Jedes Dokument wird dabei nur einmal gezählt, auch wenn der Term in ihm mehrfach vorkommt. Aus diesen Angaben lässt sich das Gewicht eines Terms bestimmen: Der Bezug zwischen Häufigkeit der Wortform und Artikellänge ergibt seine relative Häufigkeit im betrachteten Wortartikel; der Bezug zwischen Dokumentenhäufigkeit und der Gesamtanzahl an Dokumenten[5] liefert die inverse Dokumentfrequenz; das Produkt aus beiden Werten bildet das Gewicht des Terms im vorliegenden Artikel. Sortiert man nun die Terme eines jeden Artikels nach den berechneten Gewichten, so erhält man einen so genannten Relevanzvektor der Terme. Für den betrachteten Beispielartikel ist diese Sortierung in Abbildung 5 gezeigt.

[5] Beim Deutschen Wörterbuch sind dies 297.613 Dokumente.

TextID	Wort	Normierte Form	ArtikelID	Häufigkeit der Wortform	Wortanzahl im Artikel	Dokumenten-häufigkeit	Inverse Dokumentfrequenz	Gewicht
1507752	boterbloem,	boterbloem	GB13619	1	31	1	19.183	-75.854
1507766	auricomus.	auricomus	GB13619	1	31	8	16.183	-63.991
1507760	dotterblume,	dotterblume	GB13619	1	31	10	15.861	-62.718
1507749	butterblume	butterblume	GB13619	1	31	21	14.791	-58.486
1507758	caltha	caltha	GB13619	1	31	31	14.229	-56.264
1507762	leontodon	leontodon	GB13619	1	31	42	13.791	-54.531
1507763	taraxacum,	taraxacum	GB13619	1	31	53	13.455	-53.204
1507769	wähnt,	waehnt	GB13619	1	31	74	12.974	-51.300
1507765	ranunculus	ranunculus	GB13619	1	31	128	12.183	-48.174
1507759	palustris,	palustris	GB13619	1	31	163	11.834	-46.795
1507774	kühe	kuehe	GB13619	1	31	374	10.636	-42.058
1507756	kräuter,	kraeuter	GB13619	1	31	467	10.316	-40.791
1507779	gelb.	gelb	GB13619	1	31	522	10.155	-40.156
1507778	butter	butter	GB13619	1	31	666	9.804	-38.766
1507775	fressen,	fressen	GB13619	1	31	967	9.266	-36.638
1507755	mehrere	mehrere	GB13619	1	31	1408	8.724	-34.495
1507753	gilt	gilt	GB13619	1	31	2896	7.683	-30.381
1507751	nnl.	nnl	GB13619	1	31	3829	7.280	-28.788
1507768	volk	volk	GB13619	1	31	4258	7.127	-28.182
1507776	werde	werde	GB13619	1	31	5146	6.854	-27.101
TextID	Wort	Normierte Form	ArtikelID	Häufigkeit der Wortform	Wortanzahl im Artikel	Dokumenten-häufigkeit	Inverse Dokumentfrequenz	Gewicht
1507770	davon,	davon	GB13619	1	31	5680	6.711	-26.538
1507771	dasz	dasz	GB13619	1	31	35412	4.071	-16.098
1507772	sie	sie	GB13619	1	31	45488	3.710	-14.670
1507767	das	das	GB13619	1	31	84491	2.817	-11.137
1507750	f.	f	GB13619	1	31	93291	2.674	-10.572
1507754	für	fuer	GB13619	4	31	40359	3.882	-10.220

Abbildung 5: Sortierung der Wortformen nach den berechneten Termgewichten (= Relevanzvektor)

TextID	Wort	Normierte Form	ArtikelID	Häufigkeit der Wortform	Wortanzahl im Artikel	Dokumenten-häufigkeit	Inverse Dokumentfrequenz	Gewicht
1117899	Gackelblumm	gackelblumm	PG00045	1	43	1	17.251	-76.359
1117918	Gackelsblumm	gackelsblumm	PG00045	1	43	1	17.251	-76.359
1117927	Gakkelsblumm	gakkelsblumm	PG00045	1	43	1	17.251	-76.359
1117930	Kuhschelle.	kuhschelle	PG00045	1	43	3	15.666	-69.343
1117936	Gackel(s)-ei	gackelsei	PG00045	1	43	3	15.666	-69.343
1117893	Gackel(s)-blume	gackelsblume	PG00045	1	43	4	15.251	-67.506
1117903	Gackel1	gackel1	PG00045	1	43	4	15.251	-67.506
1117901	Godramst];	godramst	PG00045	1	43	5	14.929	-66.081
1117911	(Ranunculus)'	ranunculus	PG00045	1	43	7	14.444	-63.933
1117897	(Caltha	caltha	PG00045	1	43	8	14.251	-63.080
1117910	'Hahnenfuß	hahnenfuss	PG00045	1	43	9	14.081	-62.328
1117926	pulsatilla)',	pulsatilla	PG00045	1	43	9	14.081	-62.328
1117924	'Küchenschelle	kuechenschelle	PG00045	1	43	10	13.929	-61.655
1117914	Butterblume	butterblume	PG00045	1	43	11	13.792	-61.046
1117898	palustris)',	palustris	PG00045	1	43	12	13.666	-60.491

Abbildung 6: Relevanzvektor zum Artikel *Gackel*(*s*)*-blume* aus dem Pfälzischen Wörterbuch

Für jeden Wortartikel eines Wörterbuchs wird der Relevanzvektor nach einer festen Maximalanzahl von Elementen abgebrochen.[6] Auf diesem Weg entsteht also für jedes in das Netz zu integrierende Wörterbuch eine Liste von Relevanzvektoren.

Im darauffolgenden Schritt werden nun je zwei dieser Vektorlisten elementweise miteinander verglichen, das heißt, pro Vektorvergleich werden die in ihnen enthaltenen Wortformen gegeneinander geprüft und es wird die Anzahl der Übereinstimmungen ermittelt. Hier wird zunächst auf exakte Gleichheit getestet, womit eine Basis für die möglichen Verweise zwischen den Wortartikeln geschaffen wird. In weiteren Verfeine-

[6] Auf diese Weise werden die für die Vergleiche irrelevanten hochfrequenten Wortformen ausgeschlossen, da sie sich durch sehr niedrige Gewichtung auszeichnen.

rungen dieses Algorithmusschrittes könnten auch Informationen über die Wortformen berücksichtigt werden (zum Beispiel durch Lemmatisierung, Stemming, approximative Vergleiche etc.), um zusätzliche Übereinstimmungen zu ermitteln, die dann auch entsprechend bewertet werden können. Einen Vergleichsvektor zu obigem Beispiel zeigt Abbildung 6. Hierbei handelt es sich um den Artikel *Gackel*(*s*)*-blume* aus dem Pfälzischen Wörterbuch.

Vergleicht man die normierten Wortformen der Vektoren aus Abbildung 5 und Abbildung 6 miteinander, so findet man drei exakte Übereinstimmungen in den Formen *caltha*, *ranunculus* und *butterblume* innerhalb der ersten 20 gewichteten Terme. Über einen bestimmten Schwellenwert wird für den Gesamtalgorithmus vorgegeben, wann ein Vergleich positiv zu bewerten ist, das heißt, wann ein Verweis zwischen den verglichenen Wortartikeln etabliert werden soll. In obigem Beispiel bedeutet die Anzahl von drei Übereinstimmungen, dass eine Kante im Wortartikelgraphen vom Artikel *BUTTERBLUME* (Deutsches Wörterbuch) zum Artikel *Gackel*(*s*)*-blume* (Pfälzisches Wörterbuch) eingerichtet werden soll.

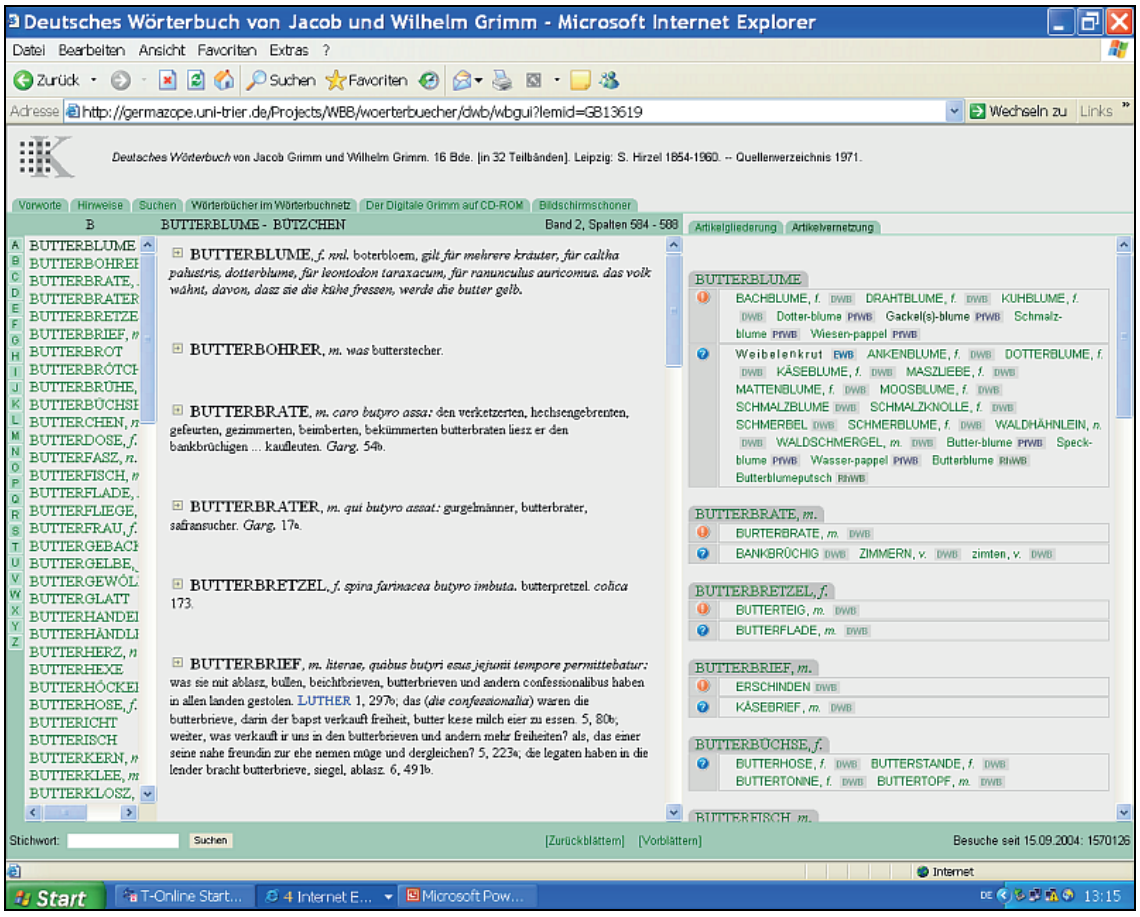

Abbildung 7: Visualisierung der Vernetzungsinformation zur Artikelstrecke ab *BUTTERBLUME* im Deutschen Wörterbuch von Jacob und Wilhelm Grimm

Jeder derart ermittelte Verweis wird in einer zusätzlichen Datenbank als Metainformation über den eigentlichen Wörterbuchdaten abgespeichert, die Basisdaten bleiben also unverändert. Beim Zugriff auf ein Wörterbuch kann aus dieser Datenbank die Vernetzungsinformation ausgelesen und ebenfalls in der grafischen Benutzeroberfläche visualisiert werden (siehe Abbildung 7).

Abbildung 7 zeigt die Vernetzung des Artikels *BUTTERBLUME* im Deutschen Wörterbuch mit weiteren Artikeln des DWB (zum Beispiel *BACHBLUME*, *DRAHTBLUME*, *KUHBLUME*, *KÄSEBLUME* etc.), aber auch mit Artikeln aus anderen Wörterbüchern (zum Beispiel *Dotterblume*, *Schmalzblume* im Pfälzischen Wörterbuch, *Weibelenkrut* im Elsässischen Wörterbuch oder *Butterblume* im Rheinischen Wörterbuch). Die Symbole vor den Verweisgruppen in Abbildung 7 kennzeichnen die Zahl der Übereinstimmungen während der Vektorenvergleiche. Eine höhere Anzahl von Übereinstimmungen führt zu einer Höhergruppierung des Verweises und bedeutet in der Regel eine höhere Zuverlässigkeit der Verknüpfung.

Das hier skizzierte ambitionierte „Idealziel", ein vielfältig verknüpftes, intelligentes „Meta-Wörterbuch" der „deutschen Sprache", steht noch in seinen Anfängen und ist gewiss nur in einer gemeinsamen, koordinierten Forschungs- und Umsetzungsanstrengung realisierbar. Die dafür notwendigen Konzepte, Methoden und Verfahren müssen in einem interdisziplinären Methodenbündel aus den folgenden Bereichen zusammenwirken: Informatik (Ontologien, Graphalgorithmen, Information-Retrieval etc.), Computerphilologie/Computerlinguistik (automatisches Markup, Pattern-Matching, Konkordanzen, automatische Lemmatisierung etc.), systematische Linguistik (Abbildung von Lautgesetzen synchron und diachron, Erarbeitung von Hyper-Lemmalisten, Phraseologie etc.), Lexikografie/Lexikologie (Semasiologie, Onomasiologie, Umkehrlexikografie). Von einem solchen Metazugriff können alle an Sprachinformationen interessierten Disziplinen in höchstem Maße profitieren, darüber hinaus sind weitreichende Impulse für neue Fragestellungen und Lösungsansätze auf verschiedensten Forschungsfeldern zu erwarten.

Literatur

Burch, Thomas/Rapp, Andrea (2007): Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven. In: Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006, hrsg. für Clio-online von Daniel Burckhardt, Rüdiger Hohls und Claudia Prinz, unter Mitwirkung von Sebastian Barteleit, Gudrun Gersmann, Peter Haber, Madeleine Herren, Patrick Sahle, Daniel Schlögl, Georg Vogeler, Claudia Wagner und Irmgard Zündorf. (= Historisches Forum 10/I). Berlin. S. 607-627. [Online-Version: http://www.akademienunion.de/projektbeschreibungen/woerterbuch-netz.htm (Stand: Mai 2008)].

Goldfarb, Charles (1990): The SGML handbook. Oxford.

Klosa, Annette/Schnörch, Ulrich/Storjohann, Petra (2006): *ELEXIKO* – A lexical and lexicological, corpus-based hypertext information system at the Institut für Deutsche Sprache, Mannheim. In: Marello, Carla et al. (Hg.): Proceedings of the 12th EURALEX International Congress (Atti del XII Congresso Internazionale di Lessicografia), EURALEX 2006, Turin, Italy, September 6th-9th, 2006. Bd. 1. Turin. S. 425-430.

Mehlhorn, Kurt (1984): Data structures and algorithms 2: Graph-algorithms and NP-completeness. (= EATCS Monographs on Theoretical Computer Science). Berlin/Heidelberg.

Müller-Spitzer, Carolin (2007a): Das *elexiko*-Portal: Ein neuer Zugang zu lexikografischen Arbeiten am Institut für Deutsche Sprache. In: Rehm, Georg/Witt, Andreas/Lemnitzer, Lothar (Hg.): Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Proceedings of the Biennial GLDV Conference 2007 (April 11-13, 2007, Eberhard-Karls-Universität Tübingen). Tübingen. S. 179-188.

Müller-Spitzer, Carolin (2007b): Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. In: Hermes 38/2007, S. 137-171.

Rapp, Andrea (2006): Das Wörterbuchnetz. Verfahren – Methoden – Perspektiven. Internet: http://www.goethe.de/wis/med/dos/dig/mew/de1526620.htm (Stand: Mai 2008).

Rob, Peter/Coronel, Carlos (1993): Database systems. Design, implementation and management. In: The Wadsworth Series in Management Information Systems. Belmont, California.

Wörterbücher

Mittelhochdeutsches Wörterbuch, Mittelhochdeutsches Handwörterbuch (einschlägige Nachträge), Findebuch zum mittelhochdeutschen Wortschatz (als Verbund mit multidirektionalen Verweisen):

BMZ (1990): Mittelhochdeutsches Wörterbuch. Mit Benutzung des Nachlasses von Georg Friedrich Benecke ausgearbeitet von Wilhelm Müller und Friedrich Zarncke. 4 Bde. Nachdruck der Ausgabe Leipzig 1854-1866. Mit einem Vorwort und einem zusammengefaßten Quellenverzeichnis von Eberhard Nellmann. Stuttgart 1990.

Burch, Thomas et al. (Hg.) (2002): Mittelhochdeutsche Wörterbücher im Verbund. CD-ROM und Begleitbuch. Stuttgart. [CD-ROM-Version auf dem CD-ROM-Server der UB Trier; Internetversion unter www.mwv.uni-trier.de].

Gärtner, Kurt et al. (1992): Findebuch zum mittelhochdeutschen Wortschatz. Datenverarbeitung: Gerhard Hanrieder. Mit einem rückläufigen Index. Stuttgart.

Lexer, Matthias (1872-1878 [1992]): Mittelhochdeutsches Handwörterbuch. 3 Bde. Nachdruck der Ausgabe Leipzig 1872-1878. Mit einer Einleitung von Kurt Gärtner. Stuttgart.

Neues Mittelhochdeutsches Wörterbuch einschließlich digitalem Quellen- und Belegarchiv:

Digitales Mittelhochdeutsches Textarchiv: http://www.mhgta.uni-trier.de.

Gärtner, Kurt et al. (Hg.) (2006): Mittelhochdeutsches Wörterbuch. Erster Bd., Doppellieferung 1/2; Lieferung 1: *a−amurschaft* bearb. in der Arbeitsstelle der Akademie der Wissenschaften und der Literatur Mainz an der Universität Trier von Ralf Plate und Jingning Tao, Lieferung 2: *an − balsieren* bearb. von der Arbeitsstelle der Akademie der Wissenschaften zu Göttingen von Susanne Baumgarte, Gerhard Diehl und Bernhard Schnell. Mit einer CD-ROM. Stuttgart. [vgl. http://www.mhdwb.uni-trier.de]

Deutsches Wörterbuch von Jacob und Wilhelm Grimm (DWB):

DWB (1984): Deutsches Wörterbuch von Jacob und Wilhelm Grimm. Nachdr. der Erstbearbeitung. München.

Der Digitale Grimm (2004): Deutsches Wörterbuch von Jacob und Wilhelm Grimm. Elektronische Ausgabe der Erstbearbeitung, bearb. von Hans-Werner Bartz, Thomas Burch, Ruth Christmann, Kurt Gärtner, Vera Hildenbrandt, Thomas Schares, Klaudia Wegge. Hrsg. vom Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier in Verbindung mit der Berlin-Brandenburgischen Akademie der Wissenschaften. 2 CD-ROMs, Benutzerhandbuch, Begleitbuch. 1. Aufl. Frankfurt a.M. Internet: http://www.dwb.uni-trier.de.

Pfälzisches Wörterbuch, Rheinisches Wörterbuch, Wörterbuch der elsässischen Mundarten, Wörterbuch der deutsch-lothringischen Mundarten, Luxemburgische Wörterbücher (Internet: http://www.dwv.uni-trier.de):

Alff, S. (Hg.)(1906): Wörterbuch der luxemburgischen Mundart. Luxemburg.

Gangler, Jean-François (1847 [2002]): Lexicon der Luxemburger Umgangssprache. Luxemburg. [Unveränderter Nachdruck Vaduz 2002].

Luxemburger Wörterbuch. Bd. 1-4 und Nachtragsband. Luxemburg 1950-1977.

Pfälzisches Wörterbuch. Begründet von Ernst Christmann, fortgeführt von Julius Krämer, bearbeitet von Rudolf Post unter Mitarbeit von Josef Schwing und Sigrid Bingenheimer. 6 Bde. Wiesbaden/ Stuttgart 1965-1997.

Rheinisches Wörterbuch. Im Auftrag der Preußischen Akademie der Wissenschaften, der Gesellschaft für Rheinische Geschichtskunde und des Provinzialverbandes der Rheinprovinz auf Grund der von Johannes Franck begonnenen, von allen Kreisen des Rheinischen Volkes unterstützten Sammlung bearbeitet und herausgegeben von Josef Müller, Heinrich Dittmaier, Rudolf Schützeichel und Mattias Zender. 9 Bde. Bonn/Berlin 1928-1971.

Wörterbuch der deutsch-lothringischen Mundarten. Bearbeitet von Ferdinand Follmann. Leipzig 1909. [Nachdruck Hildesheim/New York 1971].

Wörterbuch der elsässischen Mundarten. Bearbeitet von Ernst Martin und Hans Lienhart. 2 Bände. Straßburg 1899-1907. [Nachdruck Berlin/New York 1974].

[Vgl. hierzu auch: Fournier, Johannes (2003): Vorüberlegungen zum Aufbau eines Verbundes von Dialektwörterbüchern. In: Zeitschrift für Dialektologie und Linguistik 70, S. 155-176.]

Goethe-Wörterbuch (GWB):

GWB (1978ff.): Goethe-Wörterbuch. Hg. v. der Berlin-Brandenburgischen Akademie der Wissenschaften [bis Bd. 1, 6. Lfg.: Deutsche Akademie der Wissenschaften zu Berlin; bis Bd. 3, 4. Lfg.: Akademie der Wissenschaften der DDR], der Akademie der Wissenschaften in Göttingen und der Heidelberger Akademie der Wissenschaften. Bd. 1 (*A-azurn*), 1978; Bd. 2 (*B-einweisen*), 1989; Bd. 3 (*einwenden-Gesäusel*), 1989; Bd. 4, Lieferung IV/1-10 (*Geschäft-hinzutreten*). Internet: http://www.gwb.uni-trier.de.

Ökonomische Enzyklopädie von J. G. Krünitz:

Krünitz, Johann Georg (1773-1858): Oeconomische Encyclopädie oder allgemeines System der Land-, Haus- und Staats-Wirthschaft: in alphabetischer Ordnung. Bd. 1-242. Berlin. Internet: http://www.kruenitz.uni-trier.de.